The Occurrence of Earth-Like Planets Around Other Stars

Will M. Farr[1], Chris Aldridge[1], Kirsty Stroud[1] & Ilya Mandel[1]

[1]*School of Physics and Astronomy, University of Birmingham, Birmingham, B15 2TT, United Kingdom*

The quantity $\eta_\oplus$, the number density of planets per star per logarithmic planetary radius per logarithmic orbital period at one Earth radius and one year periods, describes the occurrence of Earth-like extrasolar planets. Measurement of $\eta_\oplus$ is complicated by the difficulty of detecting Earth-like planets in Earth-like orbits about Sun-like stars. Previous estimates[1–5] place $1\% \lesssim \eta_\oplus \lesssim 34\%$. These works dealt with the problem of selection effects in the sample by either analyzing a region of the period-radius parameter space where observations are complete and extrapolating to $R = R_\oplus$ and $P = 1\mathrm{yr}$[1,2], applying a binned analysis incorporating survey incompleteness in the period-radius plane[3,4] or analysing the results of a customised planet detection pipeline on a subset of the Kepler observations[4,5]. Here we present constraints on $\eta_\oplus$ from a parameterised forward model of the (correlated) period-radius distribution and the observational selection function in the most recent (Q17) data release from the Kepler satellite[6–8]. Our data set comprises 181,568 systems observed under the Kepler exoplanet observing program (mostly G-type stars on the main sequence[9]), producing 2598 planetary candidates. We parameterise the distribution of planetary periods and radii using a single, correlated Gaussian component; treat selection effects using a parameterised transit detection probability based on the measured noise level and stellar properties in the Kepler

catalog; and include an empirically-parameterised, independent component in the period-radius distribution to represent false-positive planet detections. Using our model we can simultaneously estimate $\eta_\oplus$, place constraints on the planet period-radius distribution function, and determine the degree of contamination by false-positive candidate identifications. We find $\eta_\oplus = 3.9^{+2.2}_{-1.6}\%$ (90% CL), in contrast to Ref. [4] but rough agreement with the analysis of the same data set in Ref. [5]. Also in contrast to Ref. [4], we can conclude that each star hosts $3.83^{+0.76}_{-0.62}$ planets with $P \lesssim 3\mathrm{yr}$ and $R \gtrsim 0.2R_\oplus$, that the peak of the planet radius distribution lies at $R_{\mathrm{peak}} = 1.25^{+0.16}_{-0.17}R_\oplus$, and that $\ln P$ and $\ln R$ are correlated with correlation coefficient $r = 0.334^{+0.052}_{-0.053}$ (all 90% CL). Our empirical model for false-positive contamination is consistent with the dominant source being background eclipsing binary stars[10], with $7.8^{+1.4}_{-1.3}\%$ (90% CL) of the candidates being false-positives.

The Kepler satellite detects planets by observing a decrement in the photometric intensity of a planet's host star as the planet transits between the telescope and the star. The Q17 data release describes 2598 "candidate" planetary transit signals identified by the Kepler team from observations of stars in the "EX" observing program (which are primarily G-type main-sequence stars similar to our own Sun[9]), giving the inferred planetary period and radius for each. The fractional depth of a planetary transit signal depends only on the radii of the planet and its host star. The signal to noise ratio of a series of transits about a particular star in the Kepler satellite scales with planetary period and radius as[11]

$$\rho = \rho_0 \left(\frac{R}{R_\oplus}\right)^2 \left(\frac{P}{1\,\mathrm{yr}}\right)^{-1/3}, \tag{1}$$

where ρ_0 is the signal to noise ratio of a Earth-radius planet in a one-year orbit about that star, which depends on the number of quarters of observation of that star, the stellar radius, and the intrinsic variability of the stellar intensity (CDPP[12]). In our analysis, we obtain these quantities from the Kepler Input Catalog[9,13] and the MAST Kepler archive[1]. To a good approximation (see Fig. 4 below), the detectability of a series of planetary transits in the Kepler data set is a function of the signal to noise ratio of the series. Because the detectability of planet transits depends on both period and radius, it is important to consider the joint (i.e. two-dimensional) distribution of these quantities in the data[14,15]. We model the detection probability as a function of signal to noise as

$$p_{\mathrm{detect}} = \begin{cases} 0 & \rho < \rho_{\mathrm{min}} \\ \frac{\log \rho - \log \rho_{\mathrm{min}}}{\log \rho_{\mathrm{max}} - \log \rho_{\mathrm{min}}} & \rho_{\mathrm{min}} < \rho < \rho_{\mathrm{max}} \\ 1 & \rho_{\mathrm{max}} < \rho \end{cases} ; \tag{2}$$

the detection probability is zero for signals with $\rho < \rho_{\mathrm{min}}$, rises linearly in $\log \rho$ for signals with $\rho_{\mathrm{min}} < \rho < \rho_{\mathrm{max}}$, and is 100% for signals with $\rho_{\mathrm{max}} < \rho$. ρ_{min} and ρ_{max} are parameters of our model. We find $\rho_{\mathrm{min}} = 5.46^{+0.18}_{-0.18}$ and $\rho_{\mathrm{max}} = 18.8^{+1.9}_{-1.9}$ (90% CL), in rough agreement with Refs. [7,8].

The probability of a planet's orbit to align with the line-of-sight to Earth and thereby produce a transit signal is

$$p_{\mathrm{transit}} = 0.0016 \, \frac{R_{\mathrm{star}}}{R_{\odot}} \left(\frac{M_{\mathrm{star}}}{M_{\odot}} \right)^{-1/3} \left(\frac{P}{1\,\mathrm{yr}} \right)^{-2/3} . \tag{3}$$

Putting Eq. 2 and 3 together, the probability that Kepler will detect a planet of radius R orbiting its host star at period P is

$$p_{\mathrm{select}} = p_{\mathrm{transit}} p_{\mathrm{detect}} . \tag{4}$$

[1] `http://archive.stsci.edu/kepler/`

We model the intrinsic distribution of planets as log-normal in period and radius, so observed planets populate the candidate P-R plane with number density

$$\frac{dN_{\mathrm{obs}}}{d\ln P\, d\ln R} = \left[\sum_{\mathrm{stars}} p_{\mathrm{select}}(P, R)\right] R_{\mathrm{pl}} N\left[\mu, \Sigma\right] \left(\ln P, \ln R\right), \tag{5}$$

where R_{pl}, μ, and Σ are parameters of our model, with R_{pl} the average number of planets per star, $\mu = \left[\mu_P, \mu_R\right]$ the mean of $\ln P$ and $\ln R$, and $\Sigma = \left[\left[\Sigma_{PP}, \Sigma_{PR}\right], \left[\Sigma_{PR}, \Sigma_{RR}\right]\right]$ the covariance matrix of $\ln P$ and $\ln R$; $N\left[\mu, \Sigma\right](x, y)$ is the normal distribution

$$N\left[\mu, \Sigma\right](x, y) = \frac{1}{2\pi\sqrt{\Sigma_{xx}\Sigma_{yy} - \Sigma_{xy}^2}} \exp\left(-\frac{1}{2}\left(\left[x, y\right] - \mu\right) \cdot \Sigma^{-1} \cdot \left(\left[x, y\right] - \mu\right)^T\right). \tag{6}$$

The expected number of observed planets in our model is

$$N_{\mathrm{pl}} = R_{\mathrm{pl}} \int dP\, dR \sum_{\mathrm{stars}} p_{\mathrm{select}}(P, R). \tag{7}$$

Our model assumes that planets appear around their host stars in a Poisson process; this is almost certainly wrong in detail[16], but nevertheless provides a good fit to the observed data (see Figure 4).

In addition to true planetary signals, we model a false-positive background of planet candidates empirically, assuming they populate the candidate P-R plane with number density

$$\frac{dN_{\mathrm{bg}}}{d\ln P\, d\ln R} = \frac{R_{\mathrm{bg}}}{\Delta \ln P \Delta \ln R}\left(1 + \vec{\gamma} \cdot \left[\ln P - \ln P_{\mathrm{mid}}, \ln R - \ln R_{\mathrm{mid}}\right]\right), \tag{8}$$

where $\Delta \ln P = \ln P_{\mathrm{max}} - \ln P_{\mathrm{min}}$, $\ln P_{\mathrm{mid}} = 1/2\left(\ln P_{\mathrm{max}} - \ln P_{\mathrm{min}}\right)$, $\Delta \ln R = \ln R_{\mathrm{max}} - \ln R_{\mathrm{min}}$, $\ln R_{\mathrm{mid}} = 1/2\left(\ln R_{\mathrm{max}} - \ln R_{\mathrm{min}}\right)$. R_{bg}, the expected number of background false-positive events; P_{max}, P_{min}, R_{max}, and R_{min}, the boundaries in the P-R plane within which background events appear; and γ, the gradient in the number density of background events, are parameters of our model.

Unlike Ref. [5], we do not attempt to model the observational uncertainties in the estimated periods and radii from the Kepler candidate data set. In spite of several candidates with very large uncertainties in measured parameters, we have found that our fit is essentially unchanged when applied to synthetic observations with periods and radii drawn re-drawn from the range of observational uncertainties quoted in the Q17 data release.

The likelihood of the observed periods and radii under our model is an inhomogeneous Poisson likelihood[15,17] with a rate that is the sum of Eq. (5) and Eq. (8). We impose priors on our 15 model parameters as follows: for the planet occurrence rate R_{pl} and (implicitly) the parameters describing selection effects, we impose a $\frac{1}{\sqrt{N_{\mathrm{pl}}}}$ prior; for the background rate R_{bg} we impose a $\frac{1}{\sqrt{R_{\mathrm{bg}}}}$ prior; for the selection model parameters ρ_{min} and ρ_{max} we impose a log-normal prior with unit width at SNRs of 3 and 11, respectively; in all other parameters we impose a flat (i.e. constant density) prior. The product of likelihood and prior gives a Bayesian posterior density function on the fifteen-dimensional parameter space of our model. We sample from this function using the `emcee` sampler of Ref. [18]. The posterior describes simultaneously the intrinsic distribution and number of exoplanets, the amount and distribution of the contaminating false-positive events in the candidate data set, and the selection function of the instrument for true planetary transit events.

The main result of this paper, the posterior distribution for $\eta_{\oplus}$, the number density of Earth-like planets, marginalised over all other parameters in our model (i.e. incorporating our uncertainty about contamination, selection effects, intrinsic distribution of planets, etc) appears in Fig. 1. Recall that

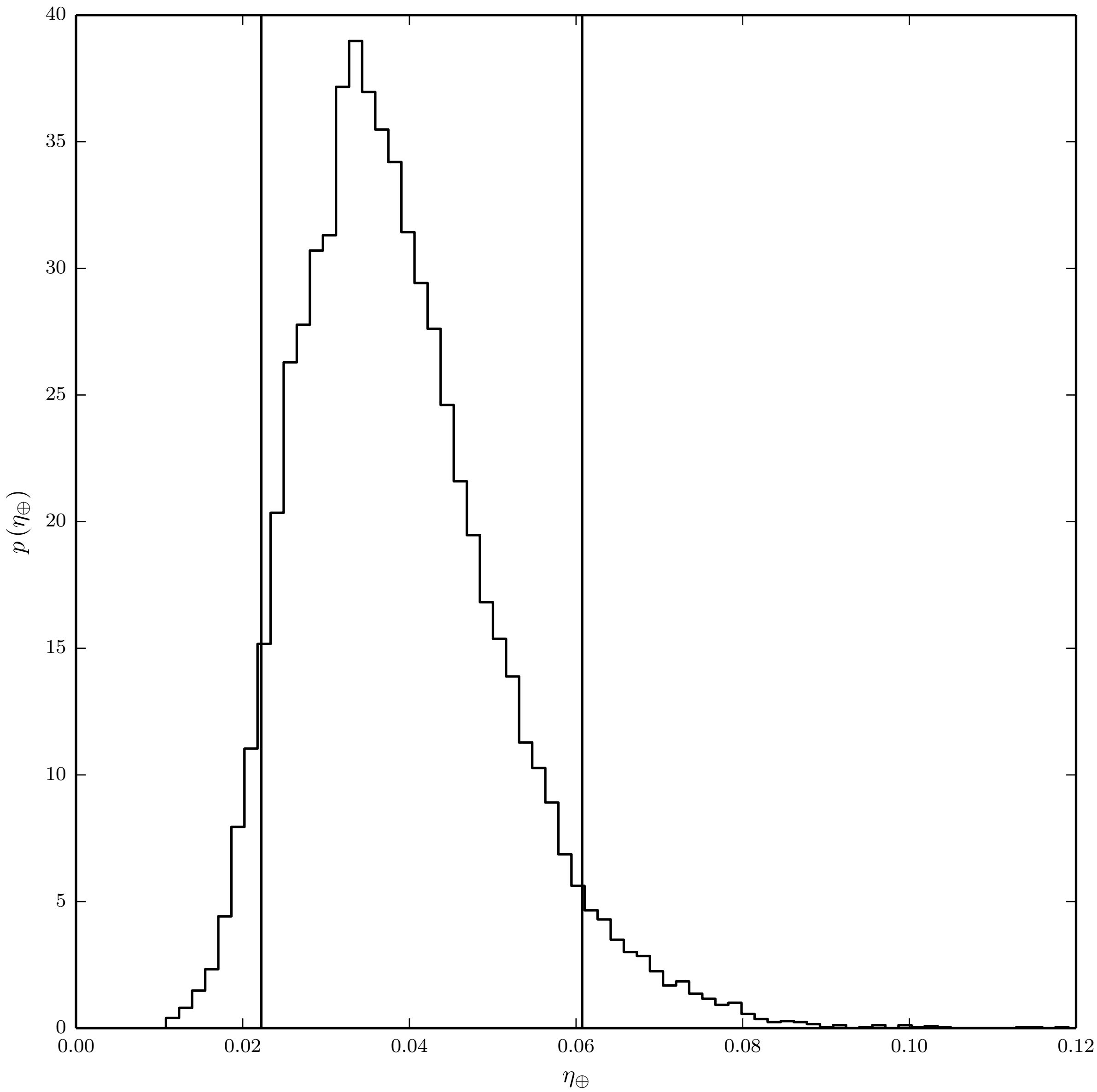
$$\eta_\oplus = \left.\frac{dN}{d\ln P \ln R}\right|_{R=R_\oplus, P=1\,\mathrm{yr}} = R_{\mathrm{pl}} N\left[\mu, \Sigma\right]\left(\ln 1\,\mathrm{yr}, \ln R_\oplus\right), \tag{9}$$

which is roughly the number of planets per star with periods and radii within a factor of $\sqrt{e}$ of Earth's. We find $\eta_\oplus = 3.9^{+2.2}_{-1.6}\%$ (90% CL). Our model also gives an estimate of the number of planets of any radius and period per star; the posterior for this quantity, marginalised over all other parameters appears in Fig. 2. We find $R_{\mathrm{pl}} = 3.83^{+0.76}_{-0.62}$ (90% CL).

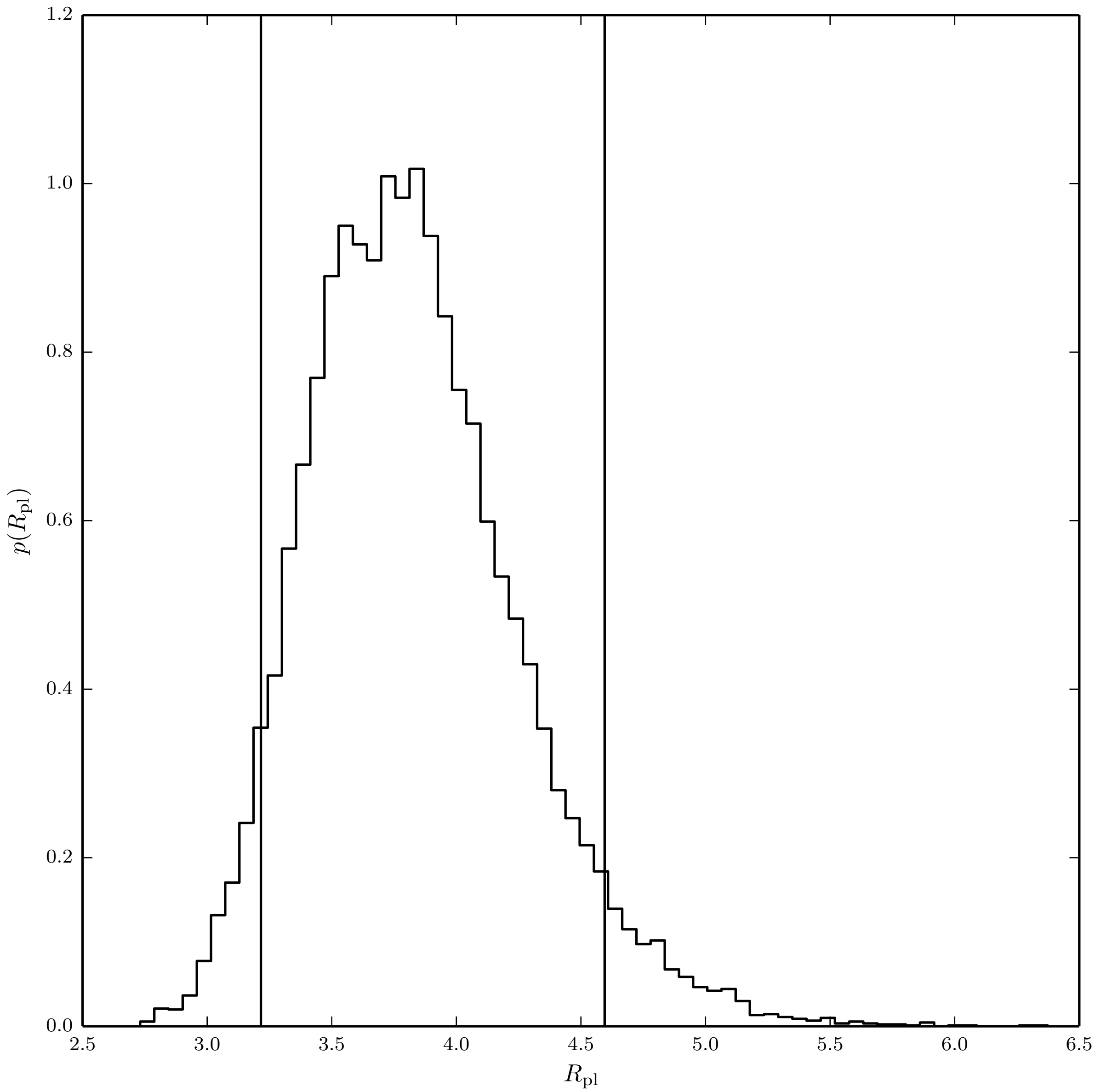
Our model also allows us to produce a posterior on the distribution of planets in the period-radius plane, and a posterior for each observed planetary candidate of the probability that this candidate is a planet instead of a background contaminant; these posteriors appear in Fig. 3. Our model finds that the false-positive rate in the candidate data set is $7.8^{+1.4}_{-1.3}\%$ (90% CL), consistent with previous work[10] estimating the contamination in the Kepler candidate set. Our model has the peak of the planet period-radius distribution at $R_{\mathrm{peak}} = 1.25^{+0.16}_{-0.17} R_\oplus$, $P_{\mathrm{peak}} = 0.075^{+0.007}_{-0.006}\mathrm{yr}$, and the distribution of planetary radii and periods is correlated, with correlation coefficient $r = 0.334^{+0.052}_{-0.053}$ (all at 90% CL).

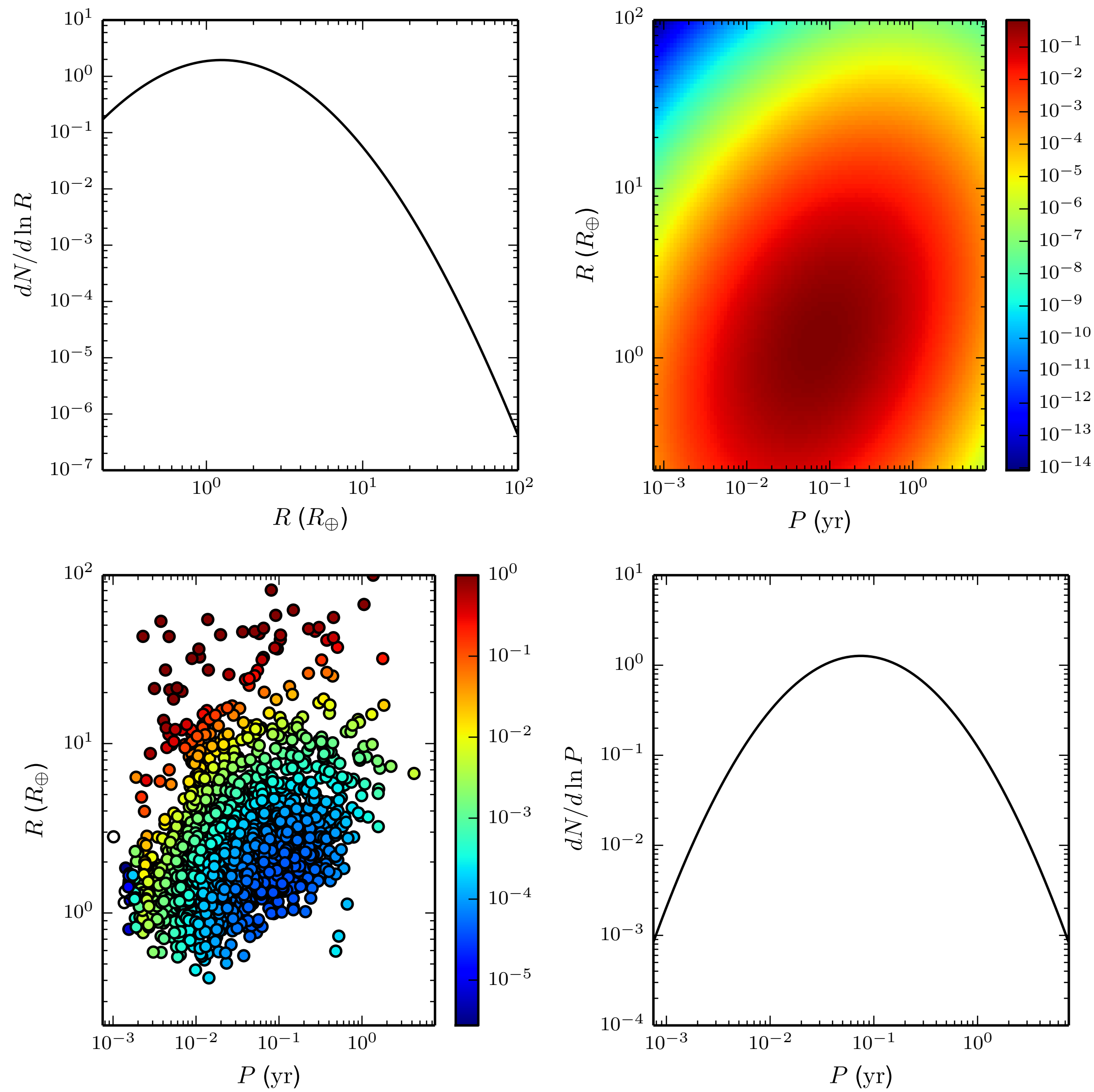
Our model predicts a distribution for future observed data consistent with the already-observed candidate set. These predictions can be used to perform graphical and posterior-predictive model checking[19]. Fig. 4 compares the predictions of our model for observed periods and radii (incorporating both planetary transits and background events) with the candidate set. This is a particularly stringent test of our parameterised selection model since the observed periods and radii are strongly influenced by the selection function of the Kepler telescope and pipeline. Except for the known sub-population of hot Jupiters[20,21], our model provides a very good fit to the observed data. That a simple log-normal distribution in period and radius fits the observed distribution of planets well may indicate that planet formation is a stochastic process with many small, correlated, and multiplicative influences on planet period and radius resulting, from the central limit theorem, in a log-normal distribution in these parameters.

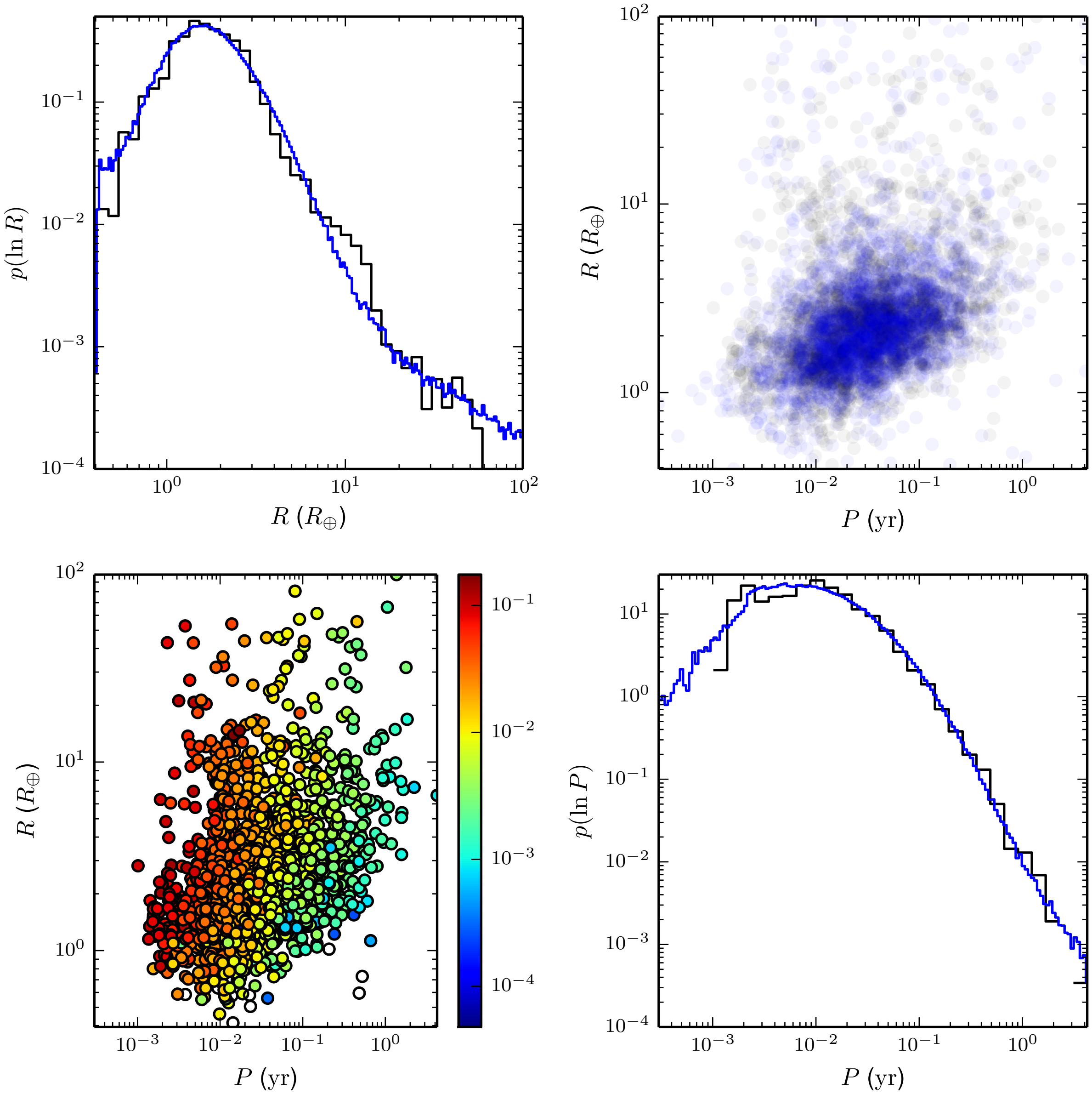
The methods and analysed data sets of Refs. [4,5] are most comparable to ours. These studies used the same data set, produced[4] from a subset of the available Kepler data and a customised pipeline to search for transit signals. They both accounted for selection effects by measuring the recoverability of synthetic transit signals injected into their data, in contrast to our approach of empirically determining them from the observed data. Neither study attempted to account for contamination from falsely-identified candidate transit events, controlling this instead through careful choice of threshold. Both studies used a more flexible model for the intrinsic distribution of planets than ours. Our result for $\eta_{\oplus}$ is consistent with, but more precise than, Ref. [5] and (somewhat) inconsistent with Ref. [4]. The reasons for the difference between these two analyses of the same data set are not clear, but our agreement with Ref. [5] lends support to that analysis and also partially validates our empirical selection model by comparison to the measured selection function in that work.



40
35
30
25
20
15
10
5
0
$p\,(\eta_{\oplus})$
0.00
0.02
0.04
0.06
0.08
0.10
0.12
$\eta_{\oplus}$


1.2
1.0
0.8
0.6
0.4
0.2
0.0
$p(R_{\rm pl})$
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
$R_{\rm pl}$


$dN/d\ln R$
$R\ (R_\oplus)$
$R\ (R_\oplus)$
P (yr)
$R\ (R_\oplus)$
P (yr)
$dN/d\ln P$
P (yr)



1. Catanzarite, J. & Shao, M. The Occurrence Rate of Earth Analog Planets Orbiting Sun-like Stars. *The Astrophysical Journal* **738**, 151 (2011). `arXiv:1103.1443`.

2. Traub, W. A. Terrestrial, Habitable-zone Exoplanet Frequency from Kepler. *The Astrophysical*

Journal **745**, 20 (2012). arXiv:1109.4682.

3. Dong, S. & Zhu, Z. Fast Rise of "Neptune-size" Planets (4-8 R $_\oplus$) from P ˜ 10 to ˜250 Days—Statistics of Kepler Planet Candidates up to ˜0.75 AU. *The Astrophysical Journal* **778**, 53 (2013). arXiv:1212.4853.

4. Petigura, E. A., Howard, A. W. & Marcy, G. W. Prevalence of Earth-size planets orbiting Sun-like stars. *Proceedings of the National Academy of Science* **110**, 19273–19278 (2013). arXiv:1311.6806.

5. Foreman-Mackey, D., Hogg, D. W. & Morton, T. D. Exoplanet population inference and the abundance of Earth analogs from noisy, incomplete catalogs. *ArXiv e-prints* (2014). arXiv:1406.3020.

6. Borucki, W. J. *et al.* Kepler Planet-Detection Mission: Introduction and First Results. *Science* **327**, 977– (2010).

7. Borucki, W. J. *et al.* Characteristics of Planetary Candidates Observed by Kepler. II. Analysis of the First Four Months of Data. *The Astrophysical Journal* **736**, 19 (2011). arXiv:1102.0541.

8. Batalha, N. M. *et al.* Planetary Candidates Observed by Kepler. III. Analysis of the First 16 Months of Data. *The Astrophysical Journal Supplement* **204**, 24 (2013). arXiv:1202.5852.

9. Batalha, N. M. *et al.* Selection, Prioritization, and Characteristics of Kepler Target Stars. *The Astrophysical Journal Letters* **713**, L109–L114 (2010). arXiv:1001.0349.

10. Fressin, F. *et al.* The False Positive Rate of Kepler and the Occurrence of Planets. *The Astrophysical Journal* **766**, 81 (2013). `arXiv:1301.0842`.

11. Chatterjee, S., Ford, E. B., Geller, A. M. & Rasio, F. A. Planets in open clusters detectable by Kepler. *Monthly Notices of the Royal Astronomical Society* **427**, 1587–1602 (2012). `arXiv:1207.3545`.

12. Christiansen, J. L. *et al.* The Derivation, Properties, and Value of Kepler"s Combined Differential Photometric Precision. *Publications of the Astronomical Society of the Pacific* **124**, 1279–1287 (2012). `arXiv:1208.0595`.

13. Brown, T. M., Latham, D. W., Everett, M. E. & Esquerdo, G. A. Kepler Input Catalog: Photometric Calibration and Stellar Classification. *The Astronomical Journal* **142**, 112 (2011). `arXiv:1102.0342`.

14. Tabachnik, S. & Tremaine, S. Maximum-likelihood method for estimating the mass and period distributions of extrasolar planets. *Monthly Notices of the Royal Astronomical Society* **335**, 151–158 (2002). `arXiv:astro-ph/0107482`.

15. Youdin, A. N. The Exoplanet Census: A General Method Applied to Kepler. *The Astrophysical Journal* **742**, 38 (2011). `arXiv:1105.1782`.

16. Weissbein, A., Steinberg, E. & Sari, R. Sterile and Fertile Planetary Systems - Statistical Analysis of Multi-Planet Systems in Kepler's data. *ArXiv e-prints* (2012). `arXiv:1203.6072`.

17. Farr, W. M., Gair, J. R., Mandel, I. & Cutler, C. Counting And Confusion: Bayesian Rate Estimation With Multiple Populations. *ArXiv e-prints* (2013). arXiv:1302.5341.

18. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific* **125**, 306–312 (2013). arXiv:1202.3665.

19. Gelman, A. *et al. Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science (Chapman & Hall/CRC, 2013), third edn.

20. Albrecht, S. *et al.* Obliquities of Hot Jupiter Host Stars: Evidence for Tidal Interactions and Primordial Misalignments. *The Astrophysical Journal* **757**, 18 (2012). arXiv:1206.6105.

21. Naoz, S., Farr, W. M. & Rasio, F. A. On the Formation of Hot Jupiters in Stellar Binaries. *The Astrophysical Journal Letters* **754**, L36 (2012). arXiv:1206.3529.

Acknowledgements The code implementing this analysis is available under an open-source "MIT" license at https://github.com/farr/kepler-selection. This work was supported by the Science and Technology Facilities Council. Computations in this work were performed on the University of Birmingham's BlueBEAR cluster. Some/all of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX13AC07G and by other grants and contracts. This paper includes data collected by the Kepler mission. Funding for the Kepler mission is provided by the NASA Science Mission directorate.

Author Contributions All authors assisted in the computational modelling, discussed the results, and edited the manuscript.

Reprints Reprints and permissions information is available at www.nature.com/reprints.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to W.M.F. (email: w.farr@bham.ac.uk).

Figure 1 The posterior on $\eta_\oplus$ accounting for selection effects and false positives. Recall that $\eta_\oplus \equiv \left.\frac{dN_{\rm pl}}{d\ln P\, d\ln R}\right|_{P=1{\rm yr}, R=1R_\oplus}$. Vertical lines indicate the 90% credible range. We find $\eta_\oplus = 3.9^{+2.2}_{-1.6}\%$, in agreement with some previous work[5], and in contrast to Ref. [4].

Figure 2 The posterior on $R_{\rm pl}$. Recall that $R_{\rm pl}$ is the number of planets per star. Vertical lines indicate the 90% credible range. We find $R_{\rm pl} = 3.83^{+0.76}_{-0.62}$.

Figure 3 The inferred planet period-radius distribution accounting for selection effects and false-positives. (Upper Left) The planet number density per logarithmic planet radius. The density peaks at $R_{\rm peak} = 1.25^{+0.16}_{-0.17} R_\oplus$ (90% CL). (Upper Right) The planet number density in the period-radius plane. The inferred correlation coefficient between $\ln P$ and $\ln R$ is $r = 0.334^{+0.052}_{-0.053}$. (Lower Left) Scatter plot of the radius and period of the Kepler planet candidates. Color indicates the posterior false-positive probability for each candidate. Overall, the model prefers a false-positive rate of $7.8^{+1.4}_{-1.3}\%$ (90% CL). The primary contaminant is probably background eclipsing binaries; our contamination rate is consistent with previous work[10]. (Lower Right) The planet number density per logarithmic planet period. The density peaks at $P = 0.075^{+0.007}_{-0.006}{\rm yr}$ (90% CL).

Figure 4 Comparison of synthetic data sets produced from the forward model incorporating selection effects with observed candidates. (Upper Left) The observed (black curve) and synthetic (blue curve; including planets and false positives, and using the fitted selection model to down-select the candidates from the planet distribution) normalised candidate density per logarithmic radius. Except for a discrepancy at $R \simeq 10R_\oplus$—associated with hot Jupiters, a distinct planetary population[20,21]—the model produces a good fit to the observed candidates over the range of reported radii. Note particularly the tail at large radii that comes from background contaminants in both observed and synthetic data. (Upper Right) Scatter plot of the observed candidates (black circles) and a posterior-averaged draw of observed candidates from the model (blue circles). (Lower Left) Scatter plot of the observed candidates. Colors indicate the posterior-averaged selection probability for each planet about its host star. Recall that the selection probability is treated a product of a geometric factor giving the probability of an isotropically-oriented orbit producing a transit and a signal-to-noise-ratio-dependent transit detection probability. (Lower Right) The observed (black curve) and synthetic (blue curve; including planets and false positives, and using the fitted selection model to down-select the candidates from the planet distribution) normalised candidate density per logarithmic period. Except for the aforementioned hot Jupiter peak at $P \simeq 1\mathrm{day}$ the model produces a good fit to the observed candidates over the range of reported periods.